1.Introduction:

1.1. Background

The fashion industry is suffering by selling something that old and people want what is on the trend. Social media platforms play a major role to change the market and the trend very quickly. Predicting what to sell next seasons is not sufficient because of social media. When customers go shopping, they have the mindset that is See-Now-Buy-Now idea and it has been affecting people all time. Meanwhile, supply chains are struggling to respond to the fashion trend. The study shows that the supply chains spent six months to put the products on the market form designing to customs on the markets. To predictive what to sell next week will be an advantage in the fashion industry. In order to sell the trend in fashion industry companies have to adopt a sales prediction model.

1.2. Problem

Costumers want the latest fashion on their hands, and it challenges the supply chains. To have the taste of costumers, the total reviews and rate of the products determine the sales of the product. The project aims to predict what products will be sold or not in the next time based on product positions.

1.3. Interest

The supply chains for the apparel industry need to respond quickly to the trend and demand. When businesses have an accurate prediction of the demand, they can cut the cost of shipping and storing.

2. Data acquisition and cleaning

2.1. data source

Analyzing historical data sales is the main source to the project. The dataset was provided by Zara. The dataset shows three-month sales for more than 14 thousand products, and what their positions on the Zara website. The dataset attributes are sales, stock, SKUs, and product positioning on the website. Table [1] shows what are the attributes and the shape of each table.

Table	Attributes	shape
products	product_id, family_id, subfamily_id, price	(15238,4)
Sales	date_number, product_id, color_id,size_id, sales,stock	(3736218,6)
Product blocks	product_id, block_id,	(15238,2)
positions	date_number, product_id, category_id, position,	(1507874,4)

Table [1] details for the dataset.

2.2. Data cleaning

The dataset is clean since the data scientists' team at Zara provides the data and big thanks for them. However, in order to have the classification for products, the data need to be combined and aggregate and drop some columns to calculate the target variables.

3. Methodology

To predict the demand for what products, have demand in the next 3 weeks, analyzing the history of the sales for 92 days is a key to solve the problems. Also, the sales history and the taste of costumers are the mina variables since the dataset has information about product positions on online stores. The mean of the product positions combined with the high positions and low positions are some features to develop the model. Besides the product positions, and a number of sales, the price for each product is the feature to progress the model for prediction. To have a sense of the features, here are visualizations for the details of the products, and product positions.

3.1. Visualizations

Form figure 1, noted the sales are high when the price is very cheap. The mean price is ($25.34) and the minimum price is ($0.33) and the maximum price is ($399.0). Most of the products very cheap because Zara is a fast-fashion brand that has a moderate price. The second feature to develop the model is the mean position for each product.

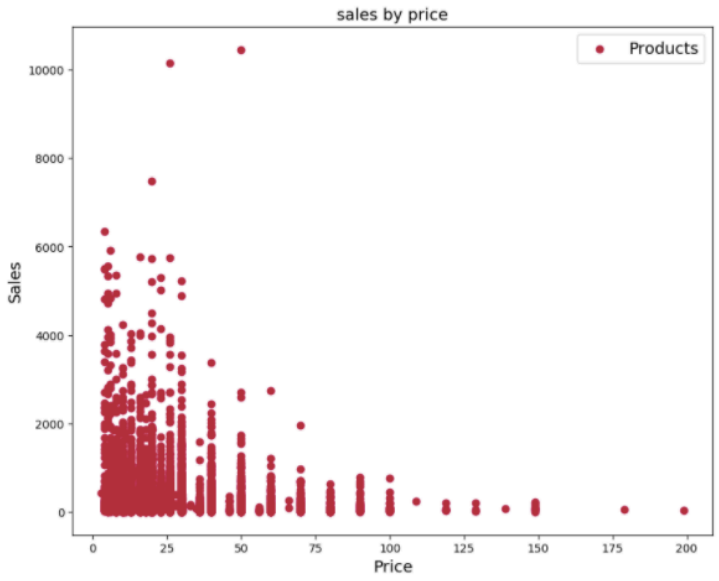

Fig. [1]

Low mean position for products has high sales since the minimum number indicates the best position for the products. For example, if the rate is 1 out of 5, the one is the best and five is the worst. With the smallest number of means, the products have a high chance to be sold in the future. Hence, the average rate of the products is a factor to be considered when the sales are predicted. Also, the products which have the best position are more likely to be sold by a high number. Figure 3 visuals how the sales are high when the products have high positions. When the costumers rate the products at a high rate, it can influence others to buy the products. Therefore, the high-rank feature was selected to develop the model. In meanwhile, including the lowest positions feature is considered progressing the model. Since the taste of costumers is different, it affects sales. Figure 4 illustrates how the sales are less when the products rated by bad positions. The chance for products to be sold is less when they have low positions. Although the figures show there are relations, they are weak relations. The function corr() in pandas.DataFrame able to calculate the correlation coefficient as the table [2] illustrates the results regression coefficient of the data frame. Moreover, the behavior of sales was vacillating in a period of 92 days and they have many trends and weak by day. Trend A and B have the highest sales by 81k and 65k respectively.

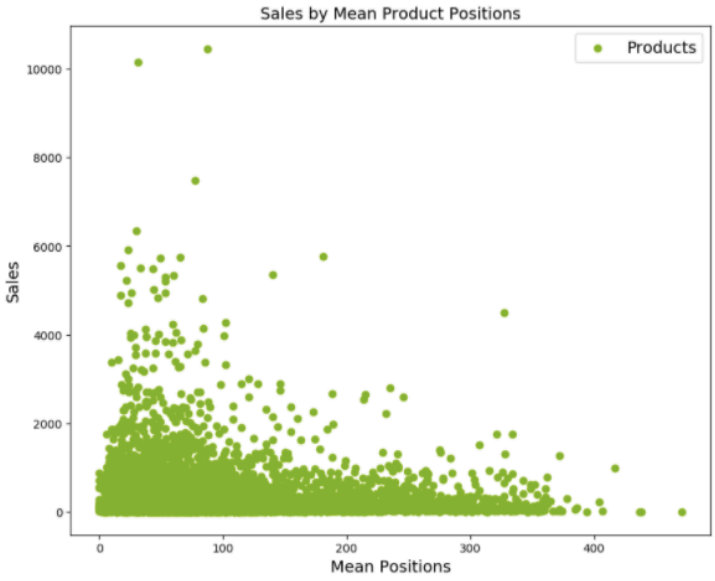

Fig. [2]

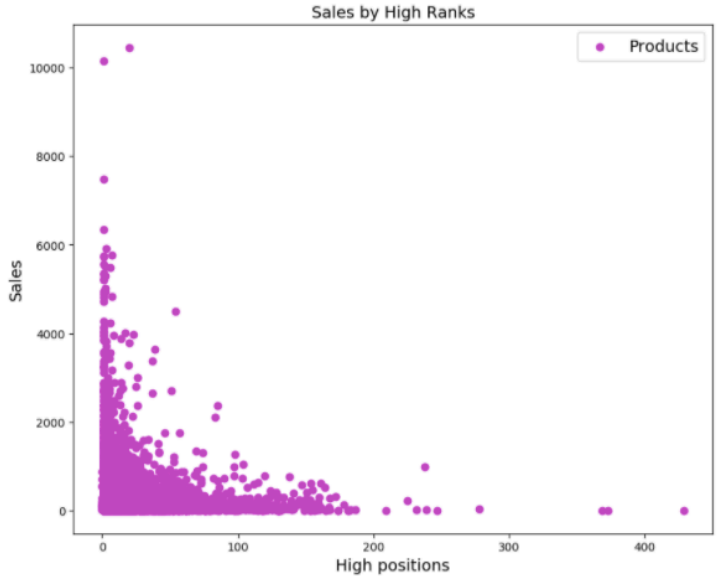

Fig. [3]

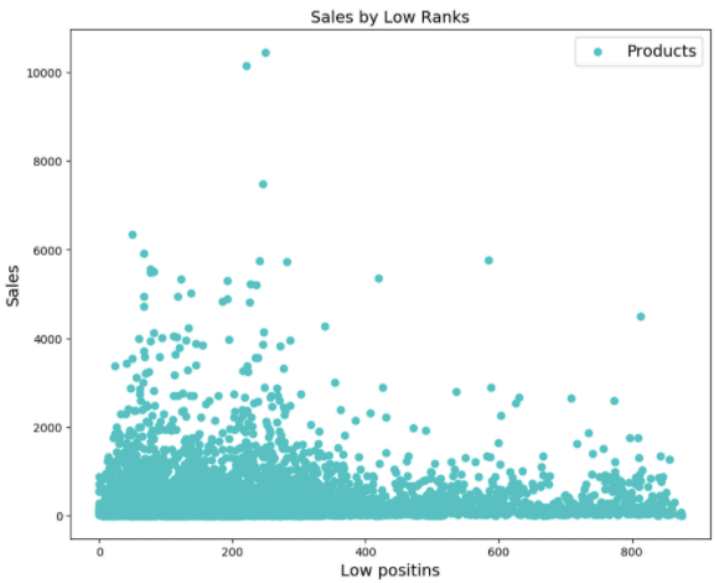

Fig: [4]

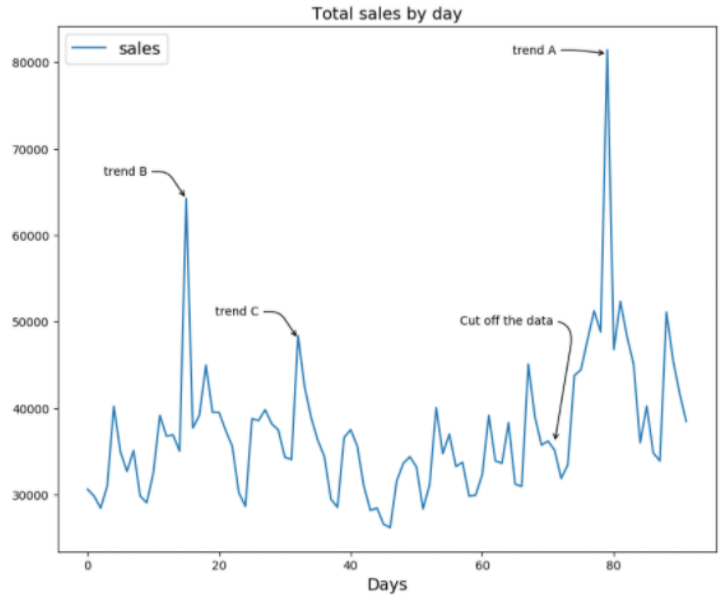

Fig: [5]

	product_id	price	sales	rank_mean	high_rank	low_rank
product_id	1.000000	-0.004423	0.007819	0.005713	0.007879	0.006618
price	-0.004423	1.000000	-0.120031	0.001335	0.046639	0.009474
sales	0.007819	-0.120031	1.000000	-0.036398	-0.083394	0.005986
rank_mean	0.005713	0.001335	-0.036398	1.000000	0.601101	0.883619
high_rank	0.007879	0.046639	-0.083394	0.601101	1.000000	0.309536
low_rank	0.006618	0.009474	0.005986	0.883619	0.309536	1.000000

Table: [2] correlation coefficient

3.2. Calculating the target variables

The data shows how the mean of product positions can affect the sales and it provides the brand of how the costumers like the products. The level of rate, in figure 6, such as 4-starts has been added to gives more sense for customers' evaluation and how they like the products. Also, the mean for positions products is the feature to label the product ranks if a product has a deal on denial. Applying cut and lambda function in python pandas to categorize the mean into two categories which are deal and denial as the figure 7 presenting. When the position products have a low mean, the product has a class deal while high mean indicates to denial position. As a result, the sales of the product are based on the product position mean which is the feature of the customers' tastes for fashion. The features which are products prince, product sales, high rank, and low rank, consider training the classification model to predict demand for the products. In other words, the models train how the mean rate of the products, then deiced if the product will be sold in the future or not.

3.3. Split dataset

The product positions and sales for each product are changing over time. Splitting the data into two sets provides the model of how the behavior of the products and sales then tested the model. The data was split into two groups called the train set and test set. The train set has a date number that less than and equal to 71-day, and the test set has a date number more than and equaled to 72-day since the project aims to predict what products demands in the next three weeks.

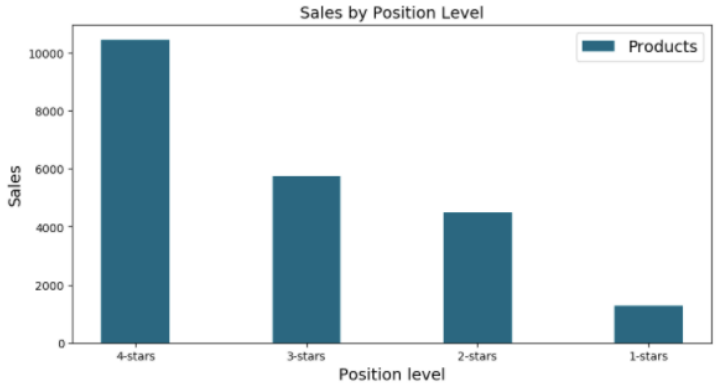

Fig: [6]

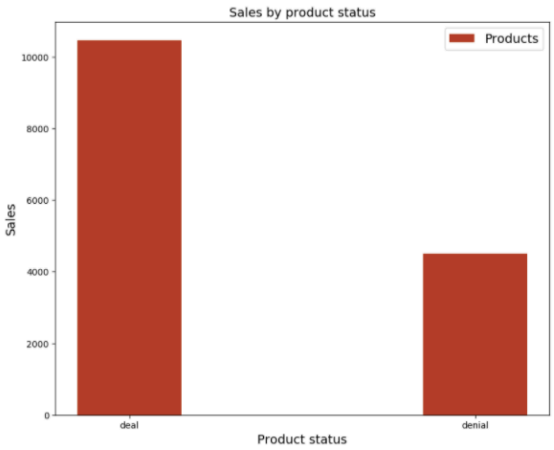

Fig: [7]

3.4. Predictive Modeling

It is clear that the regression analyses do not fit the relation and the approach is not selected to solve the problems. Since the features were selected have no liner relation showing table [2], the classification learning approach model is taken to solve the problems. The logistic regression is the statistical model was selected to develop the model since there is binary value to be predicted. The result will be if the products have a deal or denial set in the next 3 weeks. The train set including date numbers between 0 to 71, was fitting to train the model then run the model to predict the product the test set including date numbers 72 to 91.

4. Result

4.1. Prediction

The model predicts that the total number of the deal product is 8020 while 413 products are in denial set.

4.2. Model evaluation

To evaluate the model, I create a confusion matrix that allows visualization of the performance of the model. There are 387 products were Ture Negative and 261were Fales Positive. While there are 26 products were False Negative and 7759 products were Ture Positive.

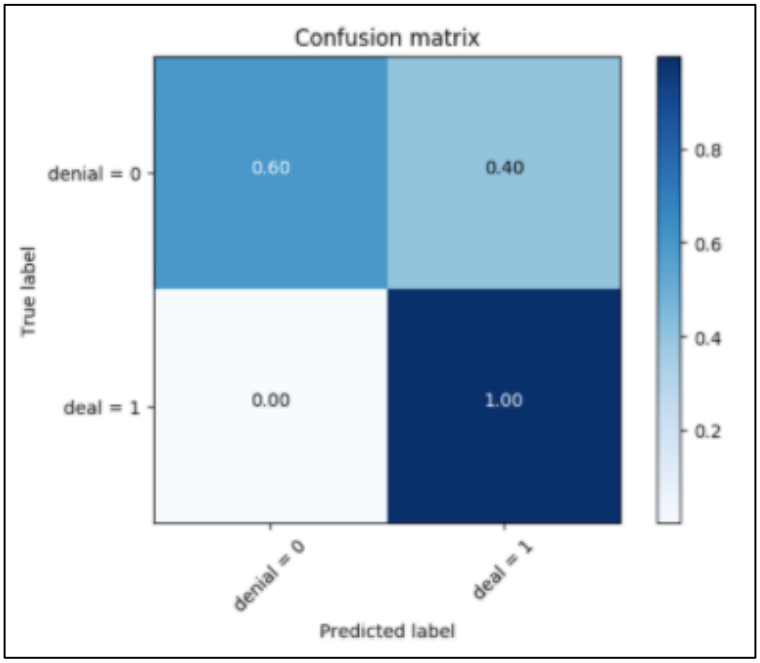

Fig: [8]

5.Disussion

Predicting demand for the fashion industry involves many factors and the most effective factors are the costumers' behaviors. The taste of costumers' changes in the short period, it is important to consider that in any forecasting model. One of the ways to explore the trend in fashion is the online stores, which may provide real-time data about how the costumers react to the products. The model provided here is not functional since the selected features make the model that bias toward the deal sets. Therefore, the classification analysis is not useful in terms of predicts the demand in fashion because it ignores many characteristics for the product such as category, color, and design. I suggest clustering analysis to consider the characteristics of products combined with product positions.

6. Conclusion

In the study, I analyzed the demand for apparel products by considering product positions. The dataset was visualized to explore relations between the features, which are total of sales, price and product positions. Also, the project included the evaluation of the model by applying the confusion matrix and what should be done in the future to have a function model for predicting the demand in the fashion industry.